Absorption Classification of Oral Drugs Based on Molecular Surface Properties

Christel A. S. Bergström,[†] Melissa Strafford,[‡] Lucia Lazorova,[†] Alex Avdeef,[‡] Kristina Luthman,*[,§] and Per Artursson[†]

Center of Pharmaceutical Informatics, Department of Pharmacy, Uppsala University, Uppsala Biomedical Center, P.O. Box 580, SE-751 23 Uppsala, Sweden, pION Inc., 5 Constitution Way, Woburn, Massachusetts 01801, and Department of Chemistry, Medicinal Chemistry, Göteborg University, SE-412 96 Göteborg, Sweden

Received July 5, 2002

The aim of this study was to investigate whether easily calculated and comprehended molecular surface properties can predict drug solubility and permeability with sufficient accuracy to allow theoretical absorption classification of drug molecules. For this purpose, structurally diverse, orally administered model drugs were selected from the World Health Organization (WHO)'s list of essential drugs. The solubility and permeability of the drugs were determined using well-established in vitro methods in highly accurate experimental settings. Descriptors for molecular surface area were generated from low-energy conformations obtained by conformational analysis using molecular mechanics calculations. Correlations between the calculated molecular surface area descriptors, on one hand, and solubility and permeability, on the other, were established with multivariate data analysis (partial least squares projection to latent structures (PLS)) using training and test sets. The obtained models were challenged with external test sets. Both solubility and permeability of the druglike molecules could be predicted with high accuracy from the calculated molecular surface properties alone. The established correlations were used to perform a theoretical biopharmaceutical classification of the WHO-listed drugs into six classes, resulting in a correct prediction for 87% of the essential drugs. An external test set consisting of Food and Drug Administration (FDA) standard compounds for biopharmaceutical classification was predicted with 77% accuracy. We conclude that PLS models of easily comprehended molecular surface properties can be used to rapidly provide absorption profiles of druglike molecules early on in drug discovery.

Introduction

Computer-based models of different complexity have been proposed as high-capacity filters in identifying poor oral absorption of druglike molecules at an early stage in drug discovery.[1,2] These models are often based on the well-established nonlinear relationship between the passive membrane permeability of drug molecules and the extent of their absorption after oral administration to humans.[3,4] However, the models do not take into account that the compounds also need sufficient water solubility in order to permeate the membrane. It has been argued that in modern drug discovery, good aqueous solubility is a more important determinant for oral drug absorption than is good membrane permeability.[1,5,6] Therefore, computer-based absorption models that take both aqueous drug solubility and permeability into account are warranted.

The realization that passive membrane permeability can be described by molecular properties made it possible to develop rapid computational models based on fairly simple molecular descriptors, such as lipophilicity, polar surface area, hydrogen bond count, and the number of rotatable bonds.[3,7−12] Surprisingly, simple molecular descriptors can also be used to provide models of aqueous drug solubility.[13−15] It should therefore be possible to combine such theoretical models in order to obtain more accurate information on the relative importance of these two parameters for oral drug absorption.

An experimental system for classification of drugs based on their aqueous solubility and membrane permeability was in fact recently implemented by the Food and Drug Administration (FDA).[16] This system, which is named the biopharmaceutics classification system (BCS), was originally implemented to waive clinical studies of generic high-permeability/high-solubility drugs. The original BCS categorizes drugs into four different classes based on combinations of high/low solubility and high/low permeability.[17] Therefore, in contrast to the rule of five,[1] which flags a potential absorption or solubility problem, a theoretical BCS would provide information about whether a compound is solubility- or permeability-limited. Thus, a computer-based BCS model would provide a more informative screening filter for the absorption properties of compound libraries in drug discovery.

Previously, we showed that the simple and easily comprehended molecular surface area descriptor polar surface area (PSA) could be used as a predictor of intestinal drug permeability.[18,19] Since then, PSA has found wide application as an absorption predictor in drug discovery.[20,21] More recently, we showed that multivariate analysis of multiple molecular surface area descriptors can be used as more accurate predictors of rather different molecular properties, such as aqueous

* To whom correspondence should be addressed. Telephone: +46 31 772 2894. Fax: +46 31 772 3840. E-mail: luthman@mc.gu.se.
[†] Uppsala University.
[‡] pION Inc.
[§] Göteborg University.

10.1021/jm020986i CCC: $25.00
Published on Web 01/11/2003

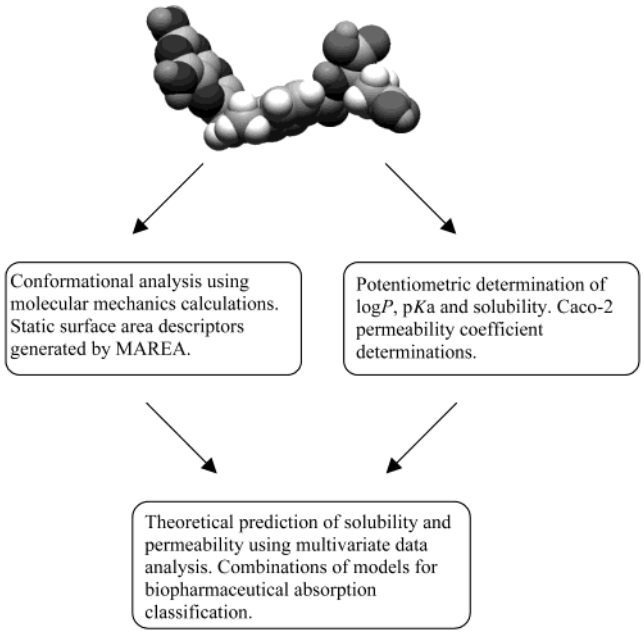

Figure 1. Flow chart of the experimental and computational studies.

solubility[22] and Caco-2 permeability.[23] We named the new surface area descriptors resulting from this analysis the "partitioned total surface areas" (PTSAs). The computer models based on PTSAs were rapidly generated and are easy to interpret. However, since separate compound datasets were used in the various studies of solubility and permeability predictions, it still remained to be investigated whether PTSA predictions hold when the same dataset is analyzed for such different absorption-related properties as solubility and permeability. We believe that only high-quality datasets should be used for this purpose, since it is likely that PTSA models established from less accurately determined experimental data for solubility and permeability will result in less predictive power.

If sufficiently accurate, PTSA modeling of solubility and permeability could be used for computer-based classification of drugs according to their absorption properties. We hypothesized that simultaneous prediction of drug solubility and permeability based on molecular surface areas would allow a simple, but more accurate, theoretical biopharmaceutical absorption classification of drugs than the rule of five[1] or similar models.[24] Such theoretical protocols should be of great value to the medicinal chemist during lead optimization (Figure 1).

Our strategy was to perform experimental determinations with high precision on smaller but relevant datasets (Figure 2), rather than in the screening mode on a larger dataset, to obtain reliable data for computer modeling. Literature data on large datasets were not considered an option because of the large interlaboratory variability in Caco-2 permeability coefficients[4] and in aqueous solubility values presented in various databases.[25] Therefore, we first experimentally determined the aqueous solubility and intestinal epithelial permeability of a series of structurally diverse drugs[26] (Figure 3), using highly accurate methods. We used the experimental data to build PTSA models for theoretical prediction of aqueous solubility and membrane permeability (Figure 4) and challenged the obtained PTSA models with external test sets obtained in-house.[22,23,27−29] Finally, these PTSA models were used for theoretical biopharmaceutical absorption classification of the drugs in our dataset as well as in an external dataset recommended by the FDA.

Results and Discussion

In this study, we have experimentally determined reliable and accurate solubility and permeability values by using the same experimental conditions for all compounds. By this strategy, we could be confident in the quality of the experimental data used in the development of predictive models for these absorption properties and for the biopharmaceutical absorption classification of drugs. Moreover, we evaluated the PTSA models by using external test sets. An important criterion for these tests was that the experimental data of the external test set should be of the same quality as the data used to build the models. An investigation of the literature shows that Caco-2 permeability determined in different laboratories can differ 10-fold in the P_{app} value.[4] The same is true for solubility values determined with different methods.[30] In our compiled external test set for solubility, where intrinsic solubility was determined using either pSOL or the small-scale shake-flask method, the R^2 of the methods was 0.96 (data not shown). Therefore, it is not realistic to assume a correlation of $R^2 = 1.0$ within a matrix consisting of compiled datasets.

Aqueous Drug Solubility. The measured solubility values ranged from 11 ng/mL (tamoxifen) to >20 mg/mL (ergonovine and zidovudine), a range of more than 6 log units (Table 1). A surprisingly good correlation ($R^2 = 0.94$, RMSE (root-mean-square error) = 0.38 log units) was obtained between the solubility values measured at 25 and 37 °C (Figure 5). These results suggest that solubility values measured in this temperature interval can be used to approximate aqueous solubility at physiological temperature (i.e., 37 °C).

A theoretical solubility model based on PTSAs was generated from the experimental solubility values presented in Table 1. Three principal components containing information mainly connected to molecular descriptors for nonpolar atoms and size were extracted (Figure 6), which resulted in an excellent model for prediction of aqueous drug solubility ($R^2 = 0.93$, $RMSE_{tr}$ (RMSE of training set) = 0.37 log units; see also Table 2). The predominant descriptors selected by the PLS analysis (Figure 6c) were those restricting solubility, which is in agreement with previous findings.[22] This supports our hypothesis that nonpolar surface areas are general molecular descriptors for aqueous drug solubility. The selection of nonpolar surface areas can be interpreted from the solvation theory:[31,32] If the nonpolar surface area of a molecule, and therefore its hydrophobicity, increases, the dissolution capacity of the compound in an aqueous environment will be decreased. Further, the selection of the size descriptor reflects the energy penalty involved in breaking the tight structure of water in order to form cavities with a large enough volume to accommodate the molecules. Only one descriptor for hydrogen bonding, the surface area of double-bonded oxygen, correlated positively with aqueous drug solubil-

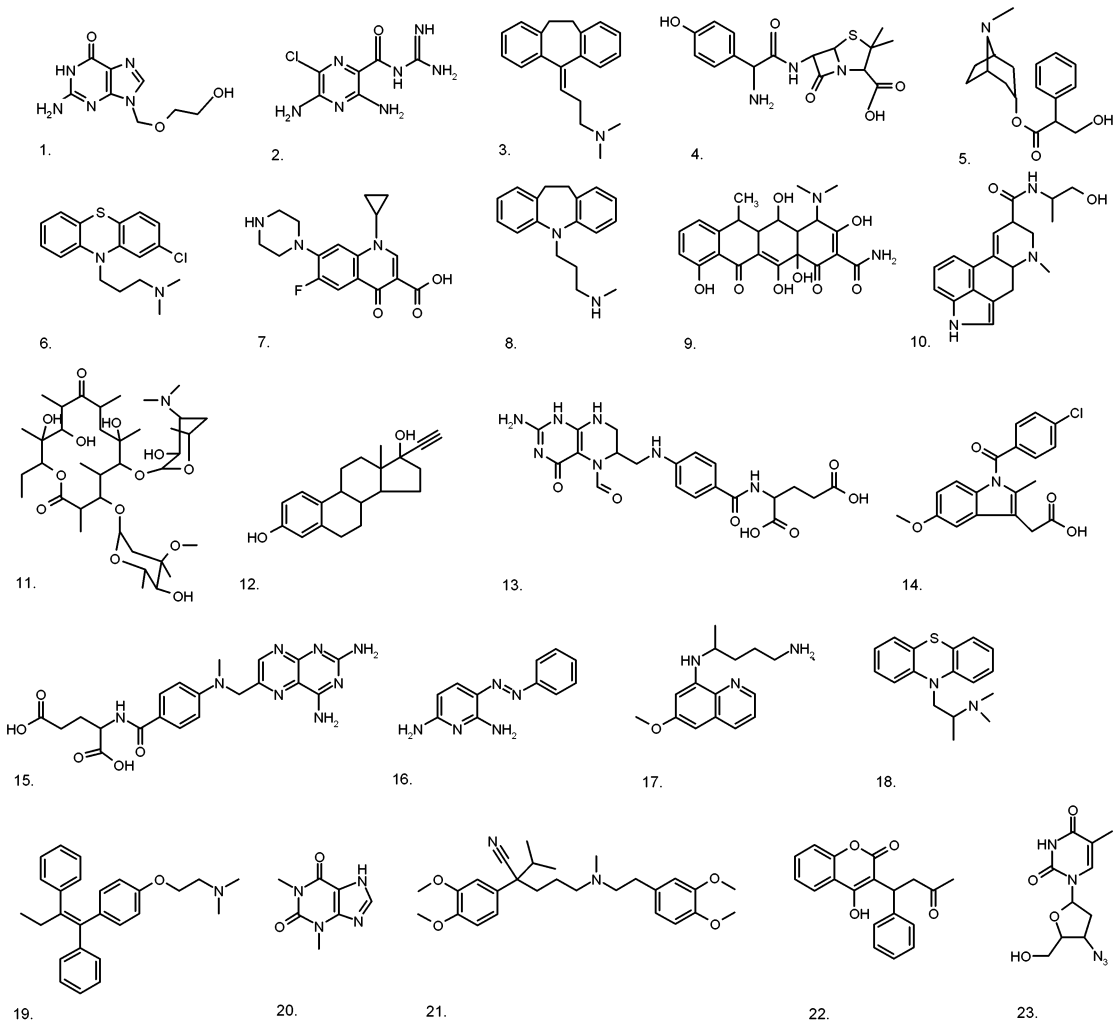

Figure 2. Chemical structures of the WHO-listed drugs used for training and testing the model: **1**, acyclovir; **2**, amiloride; **3**, amitriptyline; **4**, amoxicillin; **5**, atropine; **6**, chlorpromazine; **7**, ciprofloxacin; **8**, desipramine; **9**, doxycycline; **10**, ergonovine; **11**, erythromycin; **12**, ethinyl estradiol; **13**, folinic acid; **14**, indomethacin; **15**, methotrexate; **16**, phenazopyridine; **17**, primaquine; **18**, promethazine; **19**, tamoxifen; **20**, theophylline; **21**, verapamil; **22**, warfarin; and **23**, zidovudine.

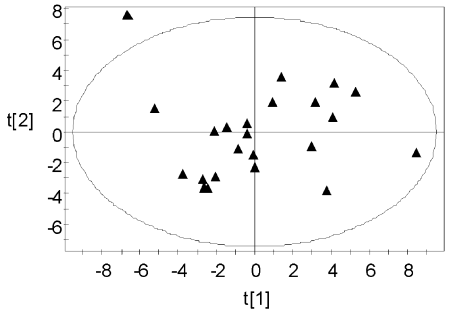

Figure 3. Heterogeneity of the selected dataset investigated by principal component analysis (PCA). The scores of the first two principal components (t1 and t2) describing 55% of the diversity in the descriptor space are shown. All molecular descriptors were used as input matrix.[66] The dataset covered all four quadrants of the PCA plot, showing that the selected series of compounds was heterogeneous. The outlier in this plot (desipramine) is well described by the other principal components extracted in the analysis. None of the 23 compounds were identified as outliers in the *x* space.

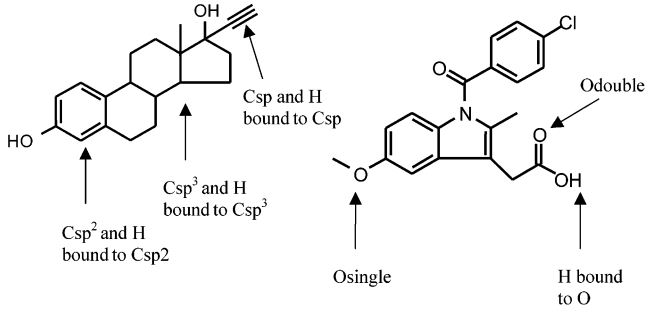

Figure 4. Examples of partitioned total surface areas (PTSAs) included in the multivariate data analysis. Nonpolar surface area (NPSA) originating from carbon atoms (Csp, Csp^2, Csp^3, and hydrogen atoms bound to carbon atoms) and polar surface area (PSA) originating from oxygen atoms (single-bonded oxygen, double-bonded oxygen, and hydrogen atoms bound to oxygen) are identified. The PTSAs represent the surface areas of each type of atom calculated with MAREA.[61]

ity. A carbonyl oxygen is a strong hydrogen bond acceptor, and the selection of this descriptor may reflect the importance of hydrogen bonding to the water molecules.

Since lipophilicity is an important descriptor included in many solubility models,[22,33,34] we expanded the solubility study with a PLS analysis of all the calculated surface properties and the lipophilicity descriptor (ClogP) to investigate whether the solubility model could be improved by incorporating a readily calculated octanol−

Table 1. Experimentally Determined Properties[a]

	compd	pK_a	log P_{oct}	−log S_0, 25 °C (M)	S class	Caco-2 ($\times 10^6$ cm/s)	a−b/b−a	P_{app} class	exptl BCS
1	acyclovir	9.23, 2.34	−1.80	2.24 ± 0.01	l	0.12 ± 0.01	0.60	l	IV
2	amiloride	8.67	−0.26	2.87 ± 0.02	h	0.32 ± 0.04	0.91	i	V
3	amitriptyline	9.49[b]	4.62	5.19 ± 0.51[b]	l	125.70 ± 7.57	1.34	h	II
4	amoxicillin	9.53, 7.31, 2.60	<−2.00	2.17 ± 0.02	l	0.18 ± 0.01	1.72	l	IV
5	atropine	9.66[b]	1.64	1.61 ± 0.10	h	19.50 ± 1.37	1.10	h	I
6	chlorpromazine	9.3[c]	5.40	5.27 ± 0.17[b]	l	72.70 ± 2.30	1.46	h	II
7	ciprofloxacin	8.66, 6.15	−1.08	3.73 ± 0.01	l	7.62 ± 2.55	0.26	h	II
8	desipramine	10.08	3.79	3.81 ± 0.04	h	101.17 ± 2.47	0.97	h	I
9	doxycycline	11.54, 8.85, 7.56, 3.21	0.52	2.35 ± 0.03	h	2.23 ± 0.21	0.46	h	I
10	ergonovine	6.91	1.62	>1.21	h	6.81 ± 0.32	1.15	h	I
11	erythromycin	8.80	2.70	3.14 ± 0.24	h	1.13 ± 0.04	0.72	i	V
12	ethinyl estradiol	10.41[b]	3.42	3.95 ± 0.12[b]	h	378.33 ± 27.74	0.87	h	I
13	folinic acid	10.15, 4.56, 3.10[b,d]	<1.50	>2.85	h	0.03 ± 0.004	1.25	l	III
14	indomethacin	4.14	3.51	5.20 ± 0.07[b]	l	109.33 ± 4.16	2.55	h	II
15	methotrexate	5.39, 4.00, 3.31[d]	0.54	4.29 ± 0.04	l	0.03 ± 0.009	1.24	l	IV
16	phenazopyridine	5.15	3.05	4.24 ± 0.02	l	284.79 ± 16.33	1.10	h	II
17	primaquine	9.99, 3.74[b]	2.72	2.77 ± 0.03	h	176.74 ± 39.99	2.17	h	I
18	promethazine	9.00[b]	4.05	4.39 ± 0.02	h	167.67 ± 20.11	1.43	h	I
19	tamoxifen	8.45	5.26	7.55 ± 0.21[b]	l	>20.00 ± 0.94[e]	0.48	h	II
20	theophylline	8.55[d]	0.00	1.38 ± 0.02	h	66.87 ± 2.31	1.30	h	I
21	verapamil	9.07[b]	4.33	4.67 ± 0.03	l	155.33 ± 17.95	1.25	h	II
22	warfarin	4.82[b]	3.54	4.74 ± 0.03	l	58.60 ± 3.96	0.73	h	II
23	zidovudine	9.53	0.13	>1.13	h	6.13 ± 0.20	0.70	h	I

[a] Lipophilicity (log P_{oct}) is given as the distribution between octanol and water. Solubility (log S_0 at 25 °C) and Caco-2 cell monolayer permeability are given as the mean ± 1 standard deviation. The ratios of transport in the absorptive (a−b) to the secretory (b−a) routes and the classification as high (h), intermediate (i), or low (l) are shown. [b] The values have been determined using a cosolvent and extrapolation to 0% (w/w) cosolvent. [c] The pK_a value was taken from Sirius Technical Application Notes (Vol 1, 1995). [d] The pK_a values were determined with the D-PAS equipment (Sirius Analytical Instruments, Forrest Row, East Sussex, U.K.). [e] Despite pretreatment of the plastics with either concentrated tamoxifen solutions or BSA, we were not successful in complete inhibition of binding of tamoxifen to plastics. Therefore, the tamoxifen permeability data were regarded as qualitative.

Table 2. Statistics of in Silico Models[a]

	solubility model	permeability model
R^2	0.93	0.93
Q^2	0.88	0.83
$RMSE_{tr}$	0.37 ($n = 14$)	0.35 ($n = 13$)
$RMSE_{te}$	0.76 ($n = 6$)	0.99 ($n = 9$)
$RMSE_{ext}$	1.05 ($n = 31$)	0.85 ($n = 26$)
$NPSA_{unsat}$	−0.81	ni
O_{dbl}	0.25	−0.32
$NPSA_{tot}$	−0.34	−0.01
$\%H_{neutral}$	0.05	0.23
SA	−0.27	−0.21
%Cl	−0.38	ni
PSA	ni	−0.31
H−N	ni	−0.25
H−O	ni	−0.13
S	ni	−0.28
constant	−2.67	−3.49

[a] Statistics from the multivariate data analysis represented by the coefficient of determination (R^2), the leave-one-out cross-validated R^2, Q^2, and the root-mean square error (RMSE) of training sets, test sets, and external test sets ($RMSE_{tr}$, $RMSE_{te}$, and $RMSE_{ext}$, respectively). The number of compounds is given in parentheses. The qualitatively measured values were not included in the RMSE calculations. The coefficients obtained from the multivariate data analysis performed in Simca are shown (ni = not included in the final model). The following descriptors were selected for description of solubility and permeability, respectively: unsaturated and total NPSA ($NPSA_{unsat}$, $NPSA_{tot}$), total surface area (SA), polar surface area (PSA), surface area of hydrogen atoms bound to nitrogen and oxygen atoms (H−N, H−O), surface area of sulfur and double-bonded oxygen atoms (S, O_{dbl}), and fraction of surface area covered by chloride atoms and electroneutral hydrogen atoms (%Cl, $\%H_{neutral}$).

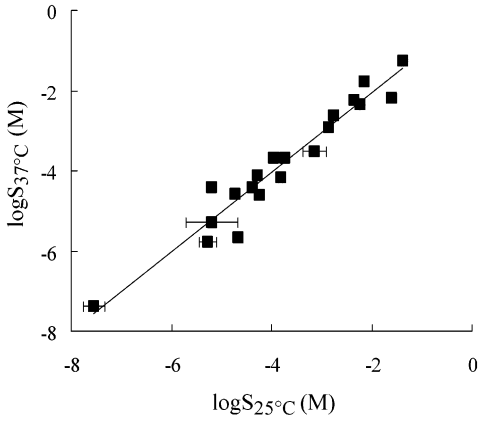

Figure 5. Correlation between solubility values determined at 25 and 37 °C resulted in an R^2 of 0.94 and an RMSE of 0.38 log units. In most cases, the standard deviation (SD) is too small to be seen in the figure. Only one of the compounds, amitriptyline, showed a large SD in its determined solubility value.

water partition coefficient. In the variable selection in the PLS analysis, the ClogP descriptor could be excluded without the model losing predictive power, which we interpreted as the molecular surface areas alone contain sufficient information regarding lipophilicity. To confirm this, we predicted the lipophilicity using the surface area descriptors selected to predict solubility (Figure 7). We found that these descriptors explained ClogP with 90% accuracy ($R^2 = 0.90$, $Q^2 = 0.84$).

The developed PTSA model for solubility was evaluated by applying an external test set of 33 compounds[22,27−29] (Figure 6d). The compounds were found to fit in the property space defined by the compounds selected from the World Health Organization (WHO)'s list,[35] since no large outliers in the x space were found (see Supporting Information). The PTSA model resulted in fair predictions ($RMSE_{ext} = 1.05$, statistical outliers probenecid and piroxicam excluded). One reason for the less accurate prediction of some of the compounds may be that the solid-state properties are not described well enough by the PTSAs obtained by the variable selection using the training set. Consequently, the conclusion of this test is that the solubility model can be used for

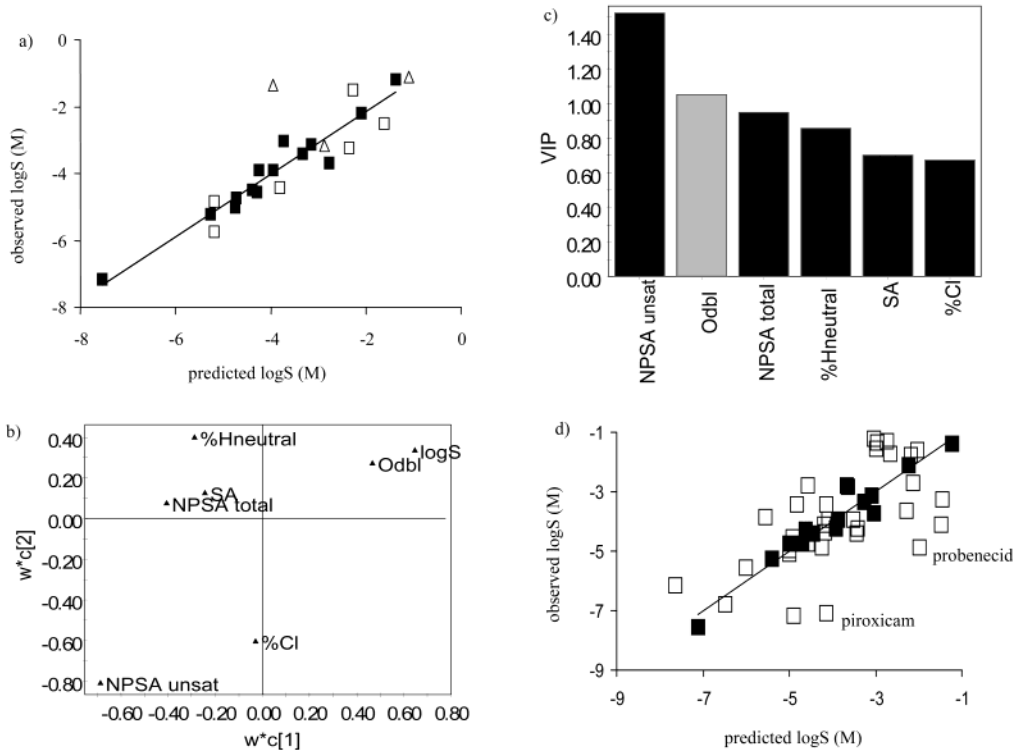

Figure 6. PLS models for in silico prediction of solubility. (a) Solubility predicted from PTSAs and composite surface areas: ■ = training set, □ = test set, and △ = qualitatively measured compounds (folinic acid, >0.7 mg/mL; ergonovine and zidovudine, >20 mg/mL). (b) Loading plot showing the interrelationship of the surface properties as descriptors of solubility. The selected composite surface areas were the following: unsaturated and total nonpolar surface area ($NPSA_{unsat}$, $NPSA_{total}$), the fraction of neutral hydrogen atoms ($\%H_{neutral}$), and the total surface area (SA). The selected PTSAs were the following: surface areas of double-bonded oxygen atoms (O_{dbl}) and fraction of chloride atoms (%Cl). (c) Variable importance on projection (VIP) plot showing the importance of each descriptor in the prediction of aqueous drug solubility. Nonpolar descriptors (shown in black) influenced the model more than did polar descriptors (shown in gray). (d) External test set[22,27−29] predicted from the PTSA model: ■ = training set, □ = external test set. Two statistical outliers (piroxicam and probenecid) were identified in the PLS prediction.

predictions of solubility but that the predictive power could be improved using an expanded training set.

At a first glance, our results seem to disagree with previous research that indicates that hydrophilic, rather than hydrophobic, descriptors are the most important molecular predictors of aqueous solubility.[14,36] This disagreement can be explained by the fact that the polar atoms in a molecule can affect drug solubility in one of two ways. Some polar atoms form energetically favorable hydrogen bonds with water, while others participate in strong intermolecular interactions within the crystal lattice. If the stabilization provided by the crystal bonds is beneficial to the molecule, the aqueous solubility will decrease, an effect described by the van't Hoff equation. Therefore, the same type of polar atom may either promote or prevent aqueous drug solubility, depending on its position in the 3D molecular structure of the compound.

Drug Permeability. The measured P_{app} coefficients ranged from 3×10^{-8} (folinic acid and methotrexate) to 4×10^{-4} cm/s (ethinyl estradiol), or more than 4 log units (Table 1). Any bias contributed by the unstirred aqueous boundary layer was minimized by efficient stirring of the Caco-2 medium. When this procedure was used, the P_{app} values could be approximated to the true cellular permeability coefficients (P_c).[23] Most compounds had a less than 2-fold difference in P_{app} in the apical to basolateral (a−b) direction compared with that in the basolateral to apical direction (b−a) and were therefore considered to be mainly passively transported (Table 1). Ciprofloxacin showed a 3.7-fold difference in the b−a direction compared with that in the a−b direction. However, a previous study in our laboratory showed that the ciprofloxacin b−a efflux contributes quantitatively only in the b−a direction, and hence, the transport rate in the absorptive (a−b) direction is concentration-independent.[23]

As for the solubility data, a theoretical model based on PTSAs was generated from the experimental P_{app} values presented in Table 1. The resulting permeability model ($R^2 = 0.93$, $RMSE_{tr} = 0.35$ log units; see also Table 2) was based on three principal components containing information mainly connected to molecular descriptors for polar atoms and size (Figure 8). The accuracy of this model was comparable to that of a previous model, generated from a different, structurally diverse dataset.[23] The negative correlation between the polar descriptors and permeability has been interpreted as a result of an increase in desolvation energy occurring when molecules enter the hydrophobic membrane interior from the aqueous surroundings.[10,37] The size of the molecule is negatively correlated to the permeability because of the steric hindrance to diffusion across the cell membrane, which is caused by the ordered membrane structure.[37] Therefore, the larger the molecule, the greater the steric hindrance to diffusion through the cell membrane.

The developed PTSA model for permeability was evaluated by applying an external test set of 27 compounds[23] (Figure 8d). For this test set, some outliers

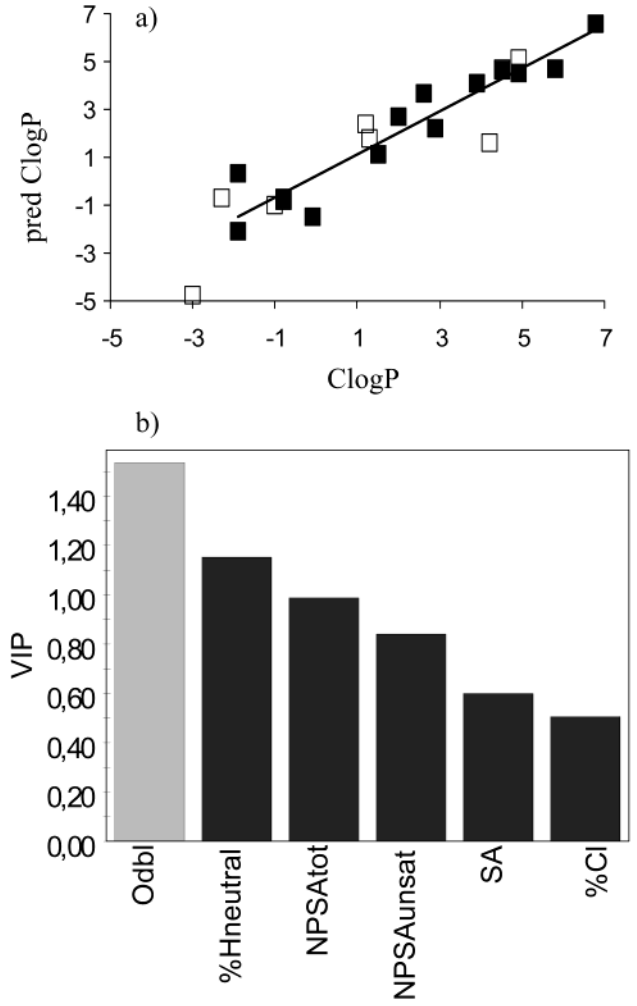

Figure 7. Molecular surface areas selected for solubility prediction as descriptors for lipophilicity (ClogP). (a) Correlation between ClogP, calculated with the ClogP program from BioByte, and ClogP predicted from the surface area of double-bonded oxygen atoms (O_{dbl}), the fraction of hydrogen bound to nonpolar atoms ($\%H_{neutral}$), the total and unsaturated nonpolar surface area ($NPSA_{tot}$, $NPSA_{unsat}$), the total surface area (SA), and the fraction of surface area of chloride atoms (%Cl). (b) Variable importance on projection (VIP) plot showing the importance of each descriptor in the prediction of ClogP. Surface properties for nonpolar and polar atoms are shown in black and gray, respectively.

were identified when fitting them to the property space defined by the compounds selected from the WHO's list (see Supporting Information).[35] However, the PTSA model resulted in good predictions, with only one statistical outlier ($RMSE_{ext} = 0.85$, mannitol excluded). We speculate that the permeability model is useful for predictions even outside the property space of the selected training set because of the robustness of used descriptors (i.e., PSA, NPSA, and SA). To summarize, by using external test sets, we showed that the selected training set used in the development of solubility and permeability models was better at predicting permeability than solubility.

Biopharmaceutical Classification. Several expressions that take both drug solubility and permeability into account have been developed.[5,38,39] One of the simplest of these is the BCS, originally proposed by the FDA. Various BCSs have previously been applied as qualitative screening tools for drug absorption in drug discovery and development.[16,40,41] In this study, we used a BCS containing six categories, according to which solubility was classified as "high" or "low"[16] and permeability was classified as "low", "intermediate", or "high".[41] In our mind, this classification provides a better tool for absorption ranking of compounds in drug discovery than does the strict permeability classification provided by the FDA.[16] Moreover, since a theoretical BCS incorporates models for both solubility and permeability, a more complete theoretical absorption profile can be obtained than from permeability predictions alone. Early information on solubility or permeability problems will strongly influence the strategy for drug development, for instance, with regard to choice of pharmaceutical dosage form. If information on these absorption characteristics is obtained and used early in drug discovery, the number of candidate drugs with formulation problems will decrease.

The experimentally determined dose-adjusted solubility values showed that 12 of the registered essential drugs (56%) were sorted as highly soluble whereas as many as 11 (44%) of the compounds showed poor solubility characteristics (Figure 9a). The compounds were found to be distributed between all three permeability classes, but only four of the compounds displayed low permeability (Table 1). Consequently, the drugs in this dataset were better optimized for permeability than for solubility. This agrees with the conclusion by Lipinski et al. that the selection of drug candidates is biased toward molecular properties giving good permeability as opposed to good solubility characteristics.[1] The combination of experimentally determined solubility and permeability data showed that the 23 compounds were distributed into five out of the six biopharmaceutical classes. As many as nine drugs (39%) in the dataset showed both high solubility and high permeability and were therefore sorted into BCS class I (Table 1, Figure 9a).

Next, we investigated whether the theoretical solubility and permeability models could be used to classify the WHO drugs,[35] using the experimental classification as a reference. To our knowledge, only one preliminary theoretical biopharmaceutical classification has been published previously.[24] First, we performed a qualitative classification in order to discriminate class I compounds from compounds with solubility and/or permeability problems. As described above, the experimental classification had identified 14 compounds that had solubility and/or permeability problems (BCS classes II−VI) (Table 1). Twelve of these (86%) were correctly identified using our PTSA models (Table 3, Figure 9b). However, two compounds (acyclovir and amitriptyline) were incorrectly classified as BCS class I compounds, that is, as false positives. In comparison, the rule of five, a commonly used theoretical screening tool for rapid assessment of permeability and solubility problems,[1] predicted that only 4 out of the 14 compounds (29%) would show solubility and/or permeability problems.

We thereafter performed a complete biopharmaceutical classification, where we sorted the compounds into classes I−VI according to their predicted solubility and permeability. Twenty out of the 23 compounds (87%) were sorted correctly into their respective class. The three compounds that were wrongly classified were

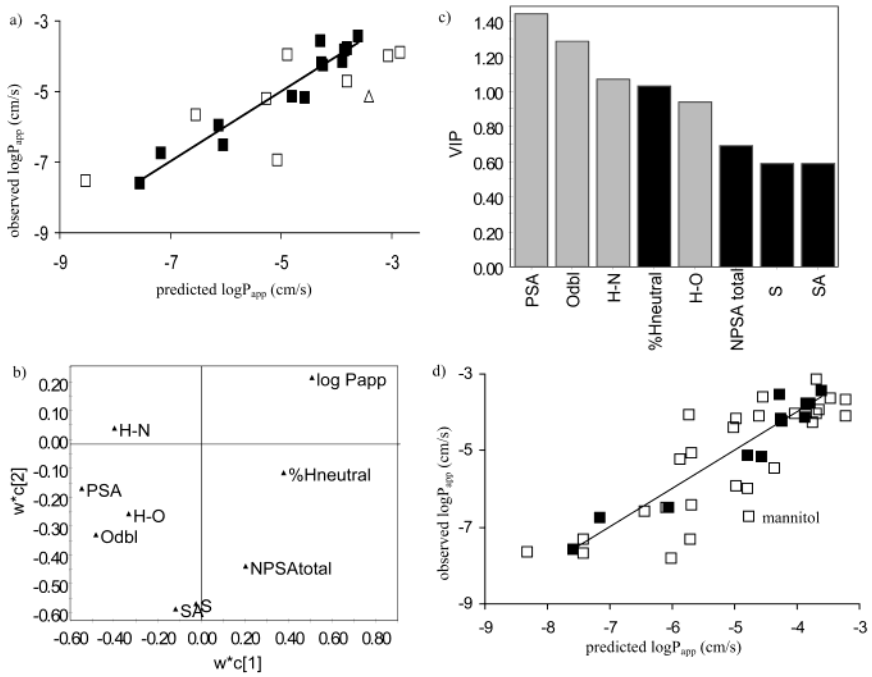

Figure 8. PLS models for in silico prediction of permeability. (a) Permeability predicted from PTSAs and composite surface areas: ■ = training set, □ = test set, and △ = qualitatively measured tamoxifen ($P_{app} > 20 \times 10^{-6}$ cm/s). (b) Loading plot showing the interrelationship of the surface properties as descriptors of drug permeability in Caco-2 cell monolayers. The selected composite surface areas were the following: polar surface area (PSA), total nonpolar surface area ($NPSA_{total}$), the fraction of neutral hydrogen atoms ($\%H_{neutral}$), and the total surface area (SA). The selected PTSAs were the following: surface areas of double-bonded oxygen (O_{dbl}), hydrogen atoms bound to nitrogen (H−N), hydrogen atoms bound to oxygen (H−O), and sulfur atoms (S). (c) Variable importance on projection (VIP) plot showing the influence of each descriptor on the permeability model. Polar surface properties (shown in gray) were more important for the model than nonpolar surface properties (shown in black). (d) External test set[23] predicted from the PTSA model: ■ = training set, □ = external test set. Only one statistical outlier, mannitol, was identified in the PLS prediction.

amitryptiline, acyclovir, and doxycycline. Amitriptyline was falsely predicted with regard to its solubility characteristics. However, amitriptyline showed a large standard deviation in its experimentally determined solubility value at 25 °C. The predicted solubility value was within the experimental error (Figure 5), providing an explanation for the erroneous prediction. The PTSA models classified acyclovir as a high-solubility/high-permeability drug, but the experimental values of both properties were low. Moreover, the theoretical values of solubility and permeability for doxycycline classified this as a low-solubility/intermediate-permeability drug, while the experimentally determined values of both properties were high. We speculate that the false predictions for acyclovir and doxycycline mean that these compounds were not correctly represented in our training set. To overcome this kind of false prediction by in silico models and to develop more generally applicable models, larger datasets covering larger parts of the structural space will be needed in the development of models.

To further evaluate the usefulness of the combinations of PTSA models in biopharmaceutical classification, we used the recommended set of reference drugs listed in the FDA guidelines.[16] This list includes 16 drugs, out of which three compounds (amoxicillin, theophylline, and verapamil) were excluded because they overlapped our dataset. The theoretical classification of this external test resulted in a correct prediction for 10 out of the 13 compounds (77%) (Table 4). Antipyrine and methyldopa were falsely predicted with regard to their solubility properties. Atenolol was predicted as a highly permeable compound but has shown low permeability in experiments. The theoretical permeability model was generated for compounds that are using the transcellular route. Atenolol has been found to use the paracellular route,[42,43] which might be the reason for the false prediction. However, none of the compounds in the external test set were falsely predicted with regard to both the solubility and the permeability characteristics.

As we have shown in the analysis of solubility, permeability, and biopharmaceutical properties, the PTSA models developed in this work have both advantages and some disadvantages. We have identified that calculated surface areas can be successfully used for prediction of two of the major properties influencing oral drug absorption, namely, solubility and permeability. Theoretical descriptors obtained from a single computational approach contain sufficient information for prediction of these absorption characteristics. Moreover, static surface areas are as successful as the dynamic surface areas when applied in these predictions, resulting in less time needed for the generation of descriptors.[23,44] In a screening mode, by use of any fairly well generated conformation, the time range for generation of descriptors would be milliseconds per structure. The obtained PTSA models have so far been built on fairly small datasets, which have been challenged with larger test sets with positive results. However, for the models to be generally applicable in the drug discovery setting, a larger structural diversity has to be investigated. The major limitation now is the generation of reliable, accurate experimental data for a large, structural

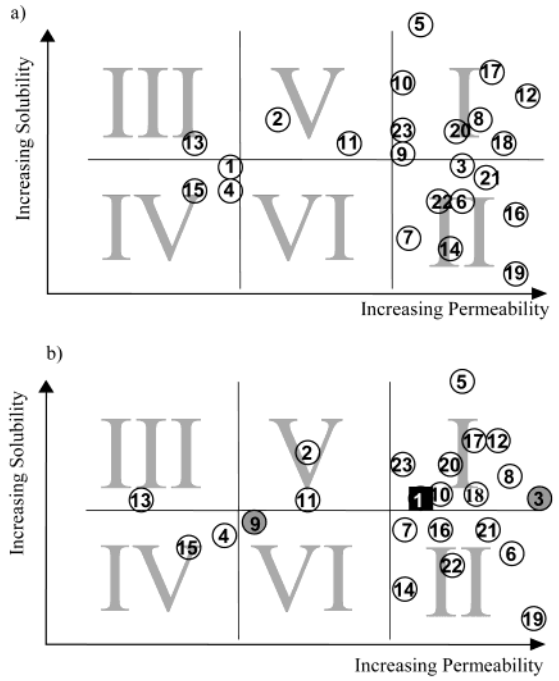

Figure 9. Comparison of experimental and theoretical biopharmaceutical classification. The six classes are marked in light-gray, and the compounds are numbered as in the tables. Relative scales are used because solubility is dose-dependent and, therefore, compound-specific. (a) Experimental determination of BCS class. The compounds were mainly distributed through classes I (39%) and II (35%). Two compounds displayed intermediate permeability, and four compounds displayed low-permeability behavior. (b) Theoretical prediction of the biopharmaceutical classes. Correctly predicted compounds are indicated with white circles, deviations from a single adjacent class are shown as light-gray circles, and shifts from a distant class are shown as black circles. The BCS classes were predicted with a success rate of 87%.

diverse dataset, which preferably should comprise several hundreds of compounds. This work will be time-consuming in any experimental setting, and therefore, such data have not yet been published. However, it is not until we obtain these kinds of data that we can address the issue of general applicability. Nevertheless, we conclude that our results are sufficiently promising to stimulate further investigations of PTSAs as rapid and transparent alternative descriptors in property-based drug design.[45]

Conclusions

In this paper, we present a new, rapid approach to theoretical biopharmaceutical classification of drug compounds, which has the capacity to significantly accelerate the absorption ranking of compound libraries in drug discovery. Our results show that multivariate data analysis of easily comprehended molecular surface descriptors provides computational tools for the prediction of both aqueous drug solubility and drug permeability. Surface areas related to the nonpolar part of the molecule resulted in good predictions of solubility, whereas surface areas describing the polar parts of the molecule resulted in good predictions of permeability. We tentatively conclude that these models will be useful for early indications regarding the absorption profiles of compound libraries at very early stages in drug discovery. Future studies based on larger datasets with reliable experimental data will further improve the general applicability of these computational protocols. Our results support the idea that simple in silico models can be assembled into more advanced physiological models that better predict oral drug absorption.[46]

Materials and Methods

Dataset. The 23 compounds used for model building were selected from the 10th revision of the World Health Organization's model list of essential drugs.[35] The selection criteria used were the following: (a) the drugs should be structurally (Figure 2), physicochemically (Figure 3), and therapeutically diverse;[47] (b) the compounds should be stable in the pH range used for the solubility titration; (c) the conformational preferences of the molecules should be possible to analyse using molecular mechanics calculations; and (d) the compounds should be mainly passively transported through the Caco-2 cell monolayers or display a concentration-independent absorption in vivo. Four of the compounds have been suggested to be actively transported (erythromycin,[48,49] verapamil,[50] folinic acid, and methotrexate[51]), although the clinical significance of the active transport in vivo is not clear for some of the compounds studied. However, these compounds were kept in the dataset to increase the structural diversity by including compounds with large size and polarity, factors that are more often associated with active rather than passive transport. It was hoped that this would result in models more useful for the drug discovery settings, where it is warranted for the investigation of large structural space and the passive transport of druglike compounds.

Except for [^{14}C]-erythromycin, which was bought from NEN Life Science Products, Inc. (Boston, MA), and [^{3}H]-tamoxifen, which was bought from Amersham Pharmacia Biotech (Uppsala, Sweden), the compounds used in this study were a gift from Charles Brownell at the FDA (MD). Amiloride, amitriptyline, chlorpromazine, desipramine, doxycycline, phenazopyridine, promethazine, and verapamil were used as their corresponding HCl salts. The maleate of ergonovine and the sulfate of atropine were used. Folinic acid was used as its corresponding calcium salt.

External Test Sets. Owing to the large variability in solubility and permeability data found in the literature,[4,25,30] our own in-house datasets were used as external test sets for the solubility and permeability models. Apparent permeability (P_{app}) values published by Stenberg and co-workers[23] were used as an external test set for permeability, while our published intrinsic solubility values[22,27−29] were compiled and used as an external test set for solubility.

The FDA recommends standard compounds for permeability classification,[16] and these compounds were used as an external test set to evaluate the obtained theoretical biopharmaceutical classification model. No corresponding list exists for solubility. The hydrophilic permeability markers listed by the FDA were excluded because they are not druglike, as were compounds already included in the dataset used to build the models. Therefore, the final external test set for the biopharmaceutical classification consisted of 13 compounds: antipyrine, atenolol, caffeine, carbamazepine, fluvastatine, furosemide, hydrochlorothiazide, ketoprofen, methyldopa, metoprolol, naproxen, ranitidine, and propranolol. For these compounds except for antipyrine,[15] caffeine,[13] carbamazepine,[52] and methyldopa,[53] we had in-house solubility data. The solubility value for ranitidine was used as a qualitative value; 5 mM solutions was used for Caco-2 experiments at pH 7.4.[54] This concentration corresponds to high solubility for ranitidine, and hence, we classified ranitidine as a highly soluble compound.

Ionization Constants, Octanol−Water Partition Coefficients, and Solubility Determinations. Prior to the solubility experiment, ionization constants (pK_a) and octanol−water partition coefficients (log P_{oct}) were determined, as described by Avdeef and co-workers.[55] The intrinsic solubility at 25 and 37 °C (±0.2 °C) was measured using pSOL Model 3 (pION Inc., Woburn, MA), as previously described.[28] Briefly,

Table 3. Calculated Properties[a]

	compd	MW	PSA[b] (Å^2)	NPSA[b] (Å^2)	predicted S class[c]	predicted P_{app} class[c]	theoretical BCS[c]
1	acyclovir*	225	128	133	h	h	I
2	amiloride	230	152	79	h	i	V
3	amitriptyline*	277	5	373	h	h	I
4	amoxicillin	365	138	260	l	l	IV
5	atropine*	289	52	309	h	h	I
6	chlorpromazine	319	8	370	l	h	II
7	ciprofloxacin	331	79	293	l	h	II
8	desipramine*	266	15	343	h	h	I
9	doxycycline*	462	165	264	l	i	VI
10	ergonovine	325	75	317	h	h	I
11	erythromycin	734	131	647	h	i	V
12	ethinyl estradiol	296	44	322	h	h	I
13	folinic acid*	473	225	259	h	l	III
14	indomethacin*	358	74	322	l	h	II
15	methotrexate	454	215	285	l	l	IV
16	phenazopyridine	213	85	178	l	h	II
17	primaquine	259	60	291	h	h	I
18	promethazine	284	5	346	h	h	I
19	tamoxifen	372	15	486	l	h	II
20	theophylline	180	74	141	h	h	I
21	verapamil	455	63	491	l	h	II
22	warfarin	308	59	301	l	h	II
23	zidovudine*	267	138	169	h	h	I

[a] Compounds marked with an asterisk (*) were selected as the test set by principal component analysis (PCA). For the solubility prediction, ergonovine maleate was added to the test set, since this compound only had a qualitative measure of solubility (>20 mg/mL). Erythromycin and tamoxifen were added to the test set of the permeability prediction, since these two compounds were analyzed differently from the others (see "Caco-2 Determinations of Permeability"). [b] Static molecular surface area descriptors: polar surface area (PSA) and nonpolar surface area (NPSA) are shown because these are the main descriptors for prediction of permeability and solubility, respectively. [c] In silico models predicting high/low (h/l) solubility class (predicted S class) and high/intermediate/low (h/i/l) permeability class (predicted P_{app} class), and the combination of these in biopharmaceutical classification (theoretical BCS) (Figure 9b).

the intrinsic solubility was determined by a pH titration of a suspension of the drug into a clear solution. The compounds were titrated in volumes of 1.7−17.0 mL and stirred with a magnetic stirrer. A stream of argon bubbles further assisted the stirring for volumes larger than 3 mL and also kept the carbonate concentration in the solution low. The results were analyzed with the pS software that accompanies pSOL, and the pK_a value and molecular weight were used to calculate intrinsic solubility.[27,28]

The solubility of ergonovine and zidovudine was >20 mg/mL. Since it was calculated that the absorption of these compounds would not be restricted by solubility and since limited amounts of the material were available, no further attempts to quantify the solubility of these two compounds were made. Moreover, the solubility of folinic acid was difficult to determine using the potentiometric technique. Although several different titration protocols were used, we were only able to determine that the solubility was >0.7 mg/mL. However, since folinic acid is administered only in low doses, it was calculated that the oral absorption of folinic acid would not be restricted by solubility. Consequently, since qualitative data were obtained for ergonovine, zidovudine, and folinic acid, they were excluded from the training set in the multivariate data analysis of solubility.

Methanol was used as cosolvent in determining the pK_a and solubility of poorly soluble compounds (Table 1). In these cases, the measurements were performed at several (three to six) different cosolvent concentrations, ranging from 4.3 to 52 w/w % methanol, and the intrinsic solubility was determined by linear extrapolation of the data to 0 w/w % methanol.[22]

Cell Culture. Caco-2 cells obtained from American Tissue Collection, Rockville, MD, were maintained in an atmosphere of 90% air and 10% CO_2, as described previously.[56] For transport experiments, 5×10^5 cells of passage number 94-100 were seeded on polycarbonate filter inserts (12 mm diameter; pore size 0.4 μm; Costar) and allowed to grow and differentiate for 21−35 days before the cell culture monolayers were used for transport experiments.

Caco-2 Determinations of Permeability. The intestinal permeability of the compounds was determined from transport rates across Caco-2 cell monolayers, as described elsewhere.[3,56] In general, the drugs were dissolved in Hank's balanced salt solution, containing 25 mM HEPES at pH 7.4 (HBSS pH 7.4), to give a final concentration of 0.02−6 mM, with each concentration being nontoxic. The amount of compound dissolved in the transport buffer depended on its solubility, its expected permeability, the presence of saturable active transport mechanisms, and its HPLC detection limit. Transport studies were initiated by incubating the monolayers in HBSS, pH 7.4, at 37 °C for 20 min in a humidified atmosphere. Filter inserts with Caco-2 cells were stirred at 500 rpm by using a plate shaker (IKA-Schüttler MTS4) during the transport experiments in order to obtain data unbiased by the aqueous boundary layer.[23] Permeability coefficients were determined both in the apical to basolateral direction and in the basolateral to apical direction (pH 7.4 in both chambers) in order to determine the possible involvement of active transport mechanisms or efflux. Monolayer permeability to the paracellular marker [^{14}C]-mannitol was routinely used to investigate the integrity of the monolayers under the experimental conditions. The samples were analyzed by HPLC, except for erythromycin (because of lack of chromophore) and tamoxifen (which was below the detection limit due to poor solubility). These two drugs were used as radioactively labeled compounds and analyzed in a liquid scintillation counter.

In two instances, the experimental protocol had to be modified in order to ensure reliable permeability data. In the first instance, erythromycin was found to be actively secreted at the applied concentration (i.e.,1.2 mM), probably by a mechanism mediated by an ABC transporter such as P-glycoprotein,[48,49] and consequently, verapamil was used to inhibit the secretion and obtain the passive permeability coefficient. The monolayers were incubated with verapamil in both the donor and the receiver chamber for 30 min prior to the transport experiments. The filters were washed, and the transport experiment was then performed in the presence of verapamil in both the apical and the basolateral chambers. This procedure using 200 μM verapamil reduced the transport in the basolateral to apical direction to the same level as the transport in the opposite direction and thus allowed the passive permeability coefficient of the Caco-2 monolayer to erythromycin to be determined. In the second instance, the

experimental protocol had to be modified because ethinyl estradiol and tamoxifen adhered to the Transwells during the transport experiments, as assessed by a mass balance calculation. The adsorption of ethinyl estradiol was prevented by saturation of the nonspecific binding sites with bovine serum albumin (BSA) (20 mg/mL) prior to the experiments. Unfortunately, the BSA only partly inhibited the plastic binding of tamoxifen. We therefore attempted to prevent the binding of tamoxifen by preincubating the filter chambers with saturated solutions of tamoxifen, but we did not succeed in obtaining complete inhibition. Consequently, the permeability of tamoxifen was determined only qualitatively to be $>2 \times 10^{-5}$ cm/s. Since it was calculated that the absorption of tamoxifen would not be restricted by permeability, no further attempts were made to quantify the permeability of this compound. As a consequence of these modified protocols, both erythromycin and tamoxifen were excluded from the training set and were instead included in the test set in the multivariate data analysis of permeability.

In general, the transport studies were performed under sink conditions and the P_{app} coefficients were calculated from

$$P_{app} = \frac{\Delta Q}{\Delta t}\frac{1}{AC_0} \quad (1)$$

where $\Delta Q/\Delta t$ is the steady-state flux (mol/s), C_0 is the initial concentration in the donor chamber at each time interval (mol/mL), and A is the surface area of the filter (cm^2). For rapidly transported compounds, where sink conditions could not be maintained for the full duration of the experiments, P_{app} was calculated, as described previously,[57] from

$$C_R(t) = \frac{M}{V_D + V_R} + \left(C_{R,0} - \frac{M}{V_D + V_R}\right) e^{-P_{app}A(1/V_D + 1/V_R)t} \quad (2)$$

where $C_R(t)$ is the time-dependent drug concentration in the receiver compartment, M is the amount of drug in the system, V_D and V_R are the volumes of the donor and receiver compartment, respectively, and t is the time from the start of the interval. P_{app} was obtained from nonlinear regression, minimizing the sum of squared residuals ($\sum(C_{R,i,obs} - C_{R,i,calc})^2$), where $C_{R,i,obs}$ is the observed receiver concentration at the end of the interval and $C_{R,i,calc}$ is the corresponding concentration calculated according to eq 2.[57]

Analytical Methods. A reversed-phase HPLC system was used to determine the drug concentration in Caco-2 samples. The HPLC system consisted of the following components: two Bischoff HPLC compact pumps, model 2250, a Bischoff LC-CaDI 22-14 integrator, a Bischoff DAD 3L-EU/3L-OU UV detector (Bischoff Analysentechnik und -geräte GmbH, Leonberg, Germany), a JASCO FP-1520 fluorescence detector (Jasco Corp., Tokyo, Japan), a Midas model 830 autosampler (Spark, Emmen, The Netherlands), and the McDAcq32 chromatography data system software, version 1.46 (Bischoff Analysentechnik und -geräte GmbH, Leonberg, Germany). A C8 analytical column (50 mm × 5.6 mm) with a mean particle size of 5 μm was used. A mobile phase gradient composed of mobile phase A containing MQ−water/acetonitrile/TFA at ratios of 99:1:0.1 and mobile phase B containing MQ−water/acetonitrile/TFA at ratios of 1:99:0.1 were used. During one gradient cycle, the mobile phase was changed from 5% to 80% acetonitrile during 1.5 min and was thereafter lowered to 5% acetonitrile within 4 min. A flow rate of 2.0 mL/min and injection volumes of 30 μL were used during the analysis.

Radioactive samples were analyzed with a liquid scintillation counter (Packard Instruments 1900CA TRI-CARB; Canberra Instruments, Downers Grove, IL).

Biopharmaceutical Classification. The drugs were classified into six different biopharmaceutical classes according to their permeability[41] and solubility:[16] (I) high solubility/high permeability, (II) low solubility/high permeability, (III) high solubility/low permeability, (IV) low solubility/low permeability, (V) high solubility/intermediate permeability, and (VI) low solubility/intermediate permeability. A drug was regarded as a highly soluble compound if the maximum dose was soluble in 250 mL of fluid in the pH interval 1−7.5. The maximum dose found in the Physicians' Desk Reference[58] and/or in FASS[59] was compared with the minimum solubility value at a pH between 1 and 7.5. Permeability was defined as "low" if it is less than 20% and as "high" if it is greater than 80% of the given dose absorbed in humans. Drugs with fraction absorbed (FA) data between these values were defined as having intermediate permeability.[41] The P_{app} values discriminating among the three permeability classes were obtained from the correlation between drug permeability in Caco-2 cells and FA established in our laboratory.[23,60] The sigmoidal function used was

$$FA = \frac{100}{1 - \left(\frac{P_{app}}{P_{app50\%}}\right)^{\gamma}} \quad (3)$$

where γ is the slope factor. This curve was used to calculate the permeability values corresponding to FA values of 20% and 80% (see Supporting Information).[23,60]

The evaluation of the models for biopharmaceutical absorption classification, using an external test set, was performed in accordance with the limitations set by the FDA. The FDA classifies solubility and permeability as either "high" or "low".[16] We defined low permeability in the evaluation as predicted permeability coefficients of $<1.6 \times 10^{-6}$ cm/s, based on our in-house correlation between fraction absorbed and Caco-2 permeability (see Figure 2 in Supporting Information).

Table 4. Biopharmaceutical Classification of Food and Drug Administration Recommended Drugs[a]

substance	perm. FDA	theor perm.	exptl sol.	theor sol.
antipyrine	h	h	h	l
atenolol	l	h	h	h
caffeine	h	h	h	h
carbamazepine	h	h	l	l
fluvastatine	h	h	l	l
furosemide	l	l	l	l
hydrochlorothiazide	l	l	h	h
ketoprofen	h	h	l	l
methyldopa	h	h	h	l
metoprolol	h	h	h	h
naproxen	h	h	l	l
propranolol	h	h	h	h
ranitidine	l	l	h	h

[a] Experimental (exptl) and PTSA-predicted (theor) solubility (sol.) and permeability (perm.) classification. The values are given as "high (h)" or "low (l)" solubility/permeability.

Lipophilicity. Lipophilicity was calculated using the ClogP program (version 2.0) from BioByte Corp. (Claremont, CA).

Conformational Analysis. A 1000 (amiloride) to 250 000 (erythromycin) step Monte Carlo conformational analysis was carried out using the BatchMin program and the MM2 force field, as implemented in MacroModel version 6.5. The conformational analysis of zidovudine was performed using MMFF instead of MM2, since the latter does not contain all the necessary parameters. In a previous study of molecules with a conformational flexibility comparable to that of the compounds in this study, we observed that the conformational analyses performed in a vacuum or in a simulated water environment resulted in molecular surface areas of the same magnitude[23] (see also Supporting Information). Therefore, to speed up the computer calculations, the conformational analyses were only performed in vacuum with the compounds in their un-ionized state. For flexible molecules, the conformational analysis was performed by two or more Monte Carlo simulations. The global minimum conformer of each search was then used as the starting conformation for a subsequent conformational search. The resulting conformers from the searches were combined, and duplicate conformers were removed.

Molecular Surface Area Calculation. The static molecular surface areas for the global minimum conformation identified in the conformational analysis were calculated, since previous studies have shown that in the analysis of not too flexible molecules, dynamic and static surface areas correspond well.[23,44] The in-house computer program MAREA[61] was used to calculate the free surface area of each atom and the molecular volume, using the van der Waals radii, with the following results: sp- and sp^2-hybridized carbons, 1.94 Å; sp^3-hybridized carbons, 1.90 Å; oxygen, 1.74 Å; nitrogen, 1.82 Å; sulfur, 2.11 Å; chloride, 2.03 Å; electroneutral hydrogen, 1.50 Å; hydrogen bound to oxygen, 1.10 Å; and hydrogen bound to nitrogen, 1.125 Å (obtained from PCMODEL, version 4.0; see Gajewski et al.[62]). The surface areas were defined as previously described.[22,23] Briefly, composite properties, such as nonpolar surface area (NPSA) and PSA, as well as PTSA descriptors, were calculated. PSA was defined as the surface area occupied by oxygen and nitrogen and by hydrogen atoms bound to these heteroatoms, whereas NPSA was defined as the total surface area (SA) minus the PSA. PTSA descriptors correspond to the surface area of a certain type of atom. For example, the NPSA originating from carbon atoms can be partitioned into the surface areas of sp-, sp^2-, and sp^3-hybridized carbon atoms and the hydrogen atoms bound to these carbon atoms. In a similar way, the PSA originating from oxygen atoms can be partitioned into the surface areas of single-bonded oxygen, double-bonded oxygen, and hydrogen atoms bound to single-bonded oxygen atoms (Figure 4). Both the absolute surface area and the surface areas relative to the SA were calculated.

Data Analysis. Solubility and permeability values were predicted by principal component analysis (PCA)[63] and partial least-squares projection to latent structures (PLS)[64] using Simca.[65] Skewed descriptors were cubic-root-transformed prior to the multivariate data analysis to avoid their being overweighted in the models. The PCA of the input matrix with all of the calculated descriptors[66] was used to divide the compound dataset into a training set of 15 compounds and a test set of 8 compounds. The training set was selected to cover a maximum range in descriptor space. This was achieved by selecting the extreme values from the first three components of the PCA. In the solubility model, ergonovine was shifted from the training set to the test set, since its solubility was too high to determine quantitatively (i.e., >20 mg/mL). Thus, all the qualitatively determined compounds were included in the test set for solubility. In the permeability model, tamoxifen, which was only determined qualitatively because of its binding to plastics, was shifted to the test set. Also, erythromycin was shifted to the test set, since a P-glycoprotein inhibitor had to be used to obtain the permeability coefficient of passive transport. The number of PLS components computed was assessed by Q^2, the leave-one-out cross-validated R^2, using seven cross-validation rounds. Only PLS components resulting in a positive Q^2 were computed, and the number of principal components was never allowed to exceed one-third of the number of observations used in the model. The models were refined through stepwise selection of the descriptors. Initially, all descriptors[67] were included in the PLS model. After the first round, the descriptor with the least influence on the prediction was deleted, and the PLS was then repeated. If the exclusion of the least important descriptor resulted in a more predictive model (as assessed by a higher Q^2), that descriptor was permanently left out of the model. This procedure was repeated until no further improvement of the model could be achieved. The predictivity of the models was assessed by RMSE of the test set ($RMSE_{te}$) and the external test set ($RMSE_{ext}$). Compounds that had a residual between the predicted and observed value of ≥2.5 standard deviations were defined as statistical outliers.

The theoretical biopharmaceutical classification was based on a combination of the multivariate data analysis of solubility and permeability values in which the predicted solubility value was adjusted for the maximum dose given.

Acknowledgment. This work was supported by Grant No. 9478 from the Swedish Medical Research Council, the Swedish Foundation for Strategic Research, the Knut and Alice Wallenberg Foundation, the Swedish Fund for Research without Animal Experiments, and GlaxoSmithKline, PA. We thank the FDA for supplying compounds for the study. We are grateful to Dr. Anders Sokolowski for providing the analytical gradient method and to Andreas Lundquist for skillful experimental assistance.

Supporting Information Available: A scatter plot showing the relationship between surface areas calculated in water and in vacuum, a graph of Caco-2 data fitted to fraction absorbed (FA) data, the distance to model in *x* space for the external test sets used to challenge the solubility and permeability models, the input descriptors with minimum and maximum values given, the predicted permeability and solubility values for training and test sets taken from the WHO model list of essential drugs, and the CAS registry number for each compound in the absorption classification dataset. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Clark, D. E.; Pickett, S. D. Computational methods for the prediction of "drug-likeness". *Drug Discovery Today* **2000**, *5*, 49−58.

(3) Artursson, P.; Karlsson, J. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochem. Biophys. Res. Commun.* **1991**, *175*, 880−885.

(4) Artursson, P.; Palm, K.; Luthman, K. Caco-2 monolayers in experimental and theoretical predictions of drug transport. *Adv. Drug Delivery Rev.* **1996**, *22*, 67−84.

(5) Curatolo, W. Physical chemical properties of oral drug candidates in the discovery and exploratory development settings. *Pharm. Sci. Technol. Today* **1998**, *1*, 387−393.

(6) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.

(7) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Beigi, F.; Lundahl, P.; Artursson, P. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors. *J. Med. Chem.* **1998**, *41*, 5382−5392.

(8) Norinder, U.;Österberg, T. Theoretical calculation and prediction of drug transport processes using simple parameters and partial least squares projections to latent structures (PLS) statistics. The use of electrotopological state indices. *J. Pharm. Sci.* **2001**, *90*, 1076−1085.

(9) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* **1999**, *88*, 807−814.

(10) Goodwin, J. T.; Conradi, R. A.; Ho, N. F.; Burton, P. S. Physicochemical determinants of passive membrane permeability: role of solute hydrogen-bonding potential and volume. *J. Med. Chem.* **2001**, *44*, 3721−3729.

(11) van de Waterbeemd, H.; Smith, D. A.; Beaumont, K.; Walker, D. K. Property-based design: optimization of drug absorption and pharmacokinetics. *J. Med. Chem.* **2001**, *44*, 1313−1333.

(12) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(13) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

(14) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the water solubilities of crystalline compounds from their chemical structures alone. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355−1359.

(15) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.

(16) Waiver of in vivo bioavailability and bioequivalence studies for immediate-release solid oral dosage forms based on a biopharmaceutics classification system, 2002. *FDA Guidance for Industry*; Federal Drug and Food Administration: Baltimore, MD. For further information, see http://www.fda.gov/cder/guidance/index.htm.
(17) The BCS system was originally introduced by the FDA to simplify the production of generic drugs. The schedule with four classes suggests that compounds showing high solubility (in comparison to the given dose) and high permeability do not need to be studied clinically after minor changes in formulations. The four BCS classes recommended by the FDA are (I) high permeability/high solubility, (II) high permeability/low solubility, (III) low permeability/high solubility, and (IV) low permeability/low solubility.
(18) Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. Correlation of drug absorption with molecular surface properties. *J. Pharm. Sci.* **1996**, *85*, 32−39.
(19) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997**, *14*, 568−571.
(20) Kelder, J.; Grootenhuis, P. D.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514−1519.
(21) Egan, W. J.; Lauri, G. Prediction of intestinal permeability. *Adv. Drug Delivery Rev.* **2002**, *54*, 273−289.
(22) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of aqueous drug solubility. *Pharm. Res.* **2002**, 182−188.
(23) Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.* **2001**, *44*, 1927−1937.
(24) van de Waterbeemd, H. The fundamental variables of the biopharmaceutics classification system (BCS): a commentary. *Eur. J. Pharm. Sci.* **1998**, *7*, 1−3.
(25) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. Aquafac 3: Aqueous functional group activity coefficients; application to the estimation of aqueous solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(26) The dataset was selected from the WHO list of essential drugs, since there is an ongoing effort in classifying these drugs in order to facilitate generic drug development in the developing countries (Junginger, lecture in September 2001). We hope that our experimental data, which cover 26% of the WHO-listed drugs, will contribute to this classification.
(27) Avdeef, A. pH−metric solubility. 1. Solubility−pH profiles from Bjerrum plots. Gibbs buffer and pK_a in the solid state. *Pharm. Pharmacol. Commun.* **1998**, *4*, 165−178.
(28) Avdeef, A.; Berger, C. M.; Brownell, C. pH−metric solubility. 2: Correlation between the acid−base titration and the saturation shake-flask solubility−pH methods. *Pharm. Res.* **2000**, *17*, 85−89.
(29) Avdeef, A.; Berger, C. M. pH−metric solubility. 3. Dissolution titration template method for solubility determination. *Eur. J. Pharm. Sci.* **2001**, *14*, 281−291.
(30) AquaSol database compiled by S. H. Yalkowsky. For further information, see www.pharm.arizona.edu/aquasol/index.html.
(31) Hildebrand, J. Solubility. XII. Regular solutions. *J. Am. Chem. Soc.* **1929**, *51*, 66−80.
(32) Scatchard, G. Equilibria in non-electrolyte solutions in relation to the vapor pressures and densities of the components. *Chem. Rev.* **1931**, *8*, 321−333.
(33) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. The linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.
(34) Yalkowsky, S. H., Banerjee, S., Eds. *Aqueous Solubility. Methods of Estimation for Organic Compounds*; Marcel Dekker Inc.: New York, 1992.
(35) In an effort to study BCS-relevant drugs, the 10th revision of the WHO list of essential drugs was used. This has been updated, and the 11th version can be found at http://www.who.int/medicines/organization/par/edl/infedl11alpha.html.
(36) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.
(37) Marrink, S. J.; Berendsen, H. J. C. Simulation of water transport through a lipid membrane. *J. Phys. Chem.* **1994**, *98*, 4155−4168.
(38) Johnson, K. C.; Swindell, A. C. Guidance in the setting of drug particle size specifications to minimize variability in absorption. *Pharm. Res.* **1996**, *13*, 1795−1798.
(39) Sanghvi, T.; Ni, N.; Yalkowsky, S. H. A simple modified absorption potential. *Pharm. Res.* **2001**, *18*, 1794−1796.
(40) Walter, E.; Janich, S.; Roessler, B. J.; Hilfinger, J. M.; Amidon, G. L. HT29-MTX/Caco-2 cocultures as an in vitro model for the intestinal epithelium: in vitro−in vivo correlation with permeability data from rats and humans. *J. Pharm. Sci.* **1996**, *85*, 1070−1076.
(41) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernäs, H.; Karlén, A. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* **1998**, *41*, 4939−4949.
(42) Adson, A.; Burton, P. S.; Raub, T. J.; Barsuhn, C. L.; Audus, K. L.; Ho, N. F. H. Passive diffusion of weak organic electrolytes across Caco-2 cell monolayers: Uncoupling the contributions of hydrodynamic, transcellular and paracellular barriers. *J. Pharm. Sci.* **1995**, *84*, 1197−1204.
(43) Collett, A.; Sims, E.; Walker, D.; He, Y. L.; Ayrton, J.; Rowland, M.; Warhurst, G. Comparison of HT29-18-C1 and Caco-2 cell lines as models for studying intestinal paracellular drug absorption. *Pharm. Res.* **1996**, *13*, 216−221.
(44) Clark, D. E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood−brain barrier penetration. *J. Pharm. Sci.* **1999**, *88*, 815−821.
(45) Property-based drug design is supported by independent structure-specific descriptors such as size, lipophilicity, and hydrophilicity.
(46) GastroPlus. For further information, visit http://www.simulations-plus.com.
(47) The substances were selected on the basis of several criteria. The structural diversity was assessed by incorporating compounds with different ring structures and functional groups. The physicochemical diversity was primarily selected on the basis of large variations in the following properties: MW (180−734), ClogP (−3 to 6.8), PSA (5−225 Å^2), NPSA (79−647 Å^2). After the selection of dataset, the physicochemical diversity was identified with PCA analysis. The compounds were also selected to represent several therapeutic classes (n = 20).
(48) Schuetz, E. G.; Yasuda, K.; Arimori, K.; Schuetz, J. D. Human MDR1 and mouse mdr1a P-glycoprotein alter the cellular retention and disposition of erythromycin, but not of retinoic acid or benzo[*a*]pyrene. *Arch. Biochem. Biophys.* **1998**, *350*, 340−347.
(49) Takano, M.; Hasegawa, R.; Fukuda, T.; Yumoto, R.; Nagai, J.; Murakami, T. Interaction with P-glycoprotein and transport of erythromycin, midazolam and ketoconazole in Caco-2 cells. *Eur. J. Pharmacol.* **1998**, *358*, 289−294.
(50) Sandstrom, R.; Karlsson, A.; Knutson, L.; Lennernas, H. Jejunal absorption and metabolism of R/S-verapamil in humans. *Pharm. Res.* **1998**, *15*, 856−862.
(51) Tamai, I.; Tsuji, A. Carrier-mediated approaches for oral drug delivery. *Adv. Drug Delivery Rev.* **1996**, *20*, 5−32.
(52) Zerrouk, N.; Chemtob, C.; Arnaud, P.; Toscani, S.; Dugue, J. In vitro and in vivo evaluation of carbamazepine-PEG 6000 solid dispersions. *Int. J. Pharm.* **2001**, *225*, 49−62.
(53) Budavari, S., Ed. *The Merck Index*, 12th ed.; Merck & Co., Inc: Whitehouse Station, NJ, 1996.
(54) Lee, K.; Thakker, D. R. Saturable transport of H2-antagonists ranitidine and famotidine across Caco-2 cell monolayers. *J. Pharm. Sci.* **1999**, *88*, 680−687.
(55) Avdeef, A. pH−metric logP. II. Refinement of partition coefficients and ionization constants of multiprotic substances. *J. Pharm. Sci.* **1993**, *82*, 183−190.
(56) Artursson, P. Epithelial transport of drugs in cell culture. I: A model for studying the passive diffusion of drugs over intestinal absorptive (Caco-2) cells. *J. Pharm. Sci.* **1990**, *79*, 476−482.
(57) Palm, K.; Luthman, K.; Ros, J.; Grasjo, J.; Artursson, P. Effect of molecular charge on intestinal epithelial drug transport: pH-dependent transport of cationic drugs. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 435−443.
(58) *Physicians' Desk Reference*, 55th ed. Medical Economics Company: Montvale, NJ, 2000.
(59) Hedstrand, A.-G., Ed. *FASS Läkemedel i Sverige 2001* (The Swedish Counterpart to Physician's Desk); Elanders: Kungsbacka, 2001.
(60) Tavelin, S. Manuscript in preparation.
(61) The program MAREA is available upon request from the authors. The program is provided free of charge for academic users. Contact Johan Gråsjö (e-mail johan.grasjo@farmaci.uu.se).
(62) Gajewski, J. J.; Gilbert, K. E.; McKelvey, J. MMX an enhanced version of MM2. *Adv. Mol. Model.* **1990**, *2*, 65−92.
(63) Jackson, E. J. *A Users Guide to Principal Components*; Wiley: New York, 1991.
(64) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211−228.
(65) *Simca-P*, version 8.0; Umetrics AB (Box 7960, SE-907 19 Umeå, Sweden).

(66) The input matrix consisted of the following variables: MW, ClogP, Csp^2, Csp^3, O_{dbl}, O_{single}, Nsp^2, Nsp^3, $H_{neutral}$, H.-O, H.-N, S, Cl, $NPSA_{sat}$, $NPSA_{unsat}$, $NPSA_{tot}$, PSA, SA, V, $\%Csp^2$, $\%Csp^3$, $\%O_{dbl}$, $\%O_{single}$, $\%Nsp^2$, $\%Nsp^3$, $\%H_{neutral}$, %H−O, %H−N, %S, %Cl, $\%NPSA_{sat}$, $\%NPSA_{unsat}$, $\%NPSA_{tot}$, and %PSA, 1999.

(67) The PTSAs investigated in this dataset were the following: sp^2- and sp^3-hybridized carbons, sp^2-hybridized nitrogens, double- and single-bonded oxygen, sulphur, chloride, and hydrogen atoms bound to nitrogen, oxygen, and carbon atoms.

JM020986I